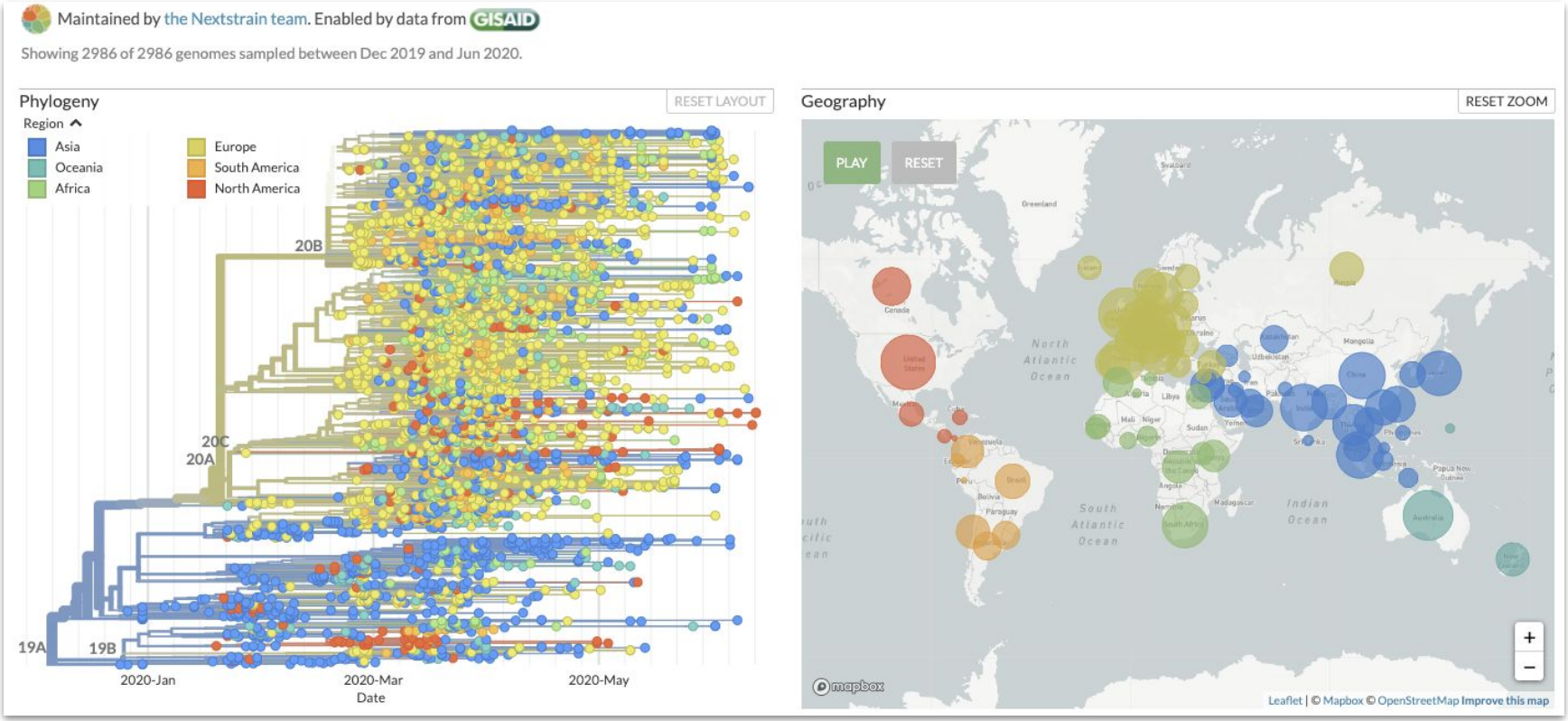
Building and analyzing SARS-CoV-2 consensus genomes

CHAN ZUCKERBERG BIOHUB

CHAN ZUCKERBERG INITIATIVE

Consensus genomes are necessary!

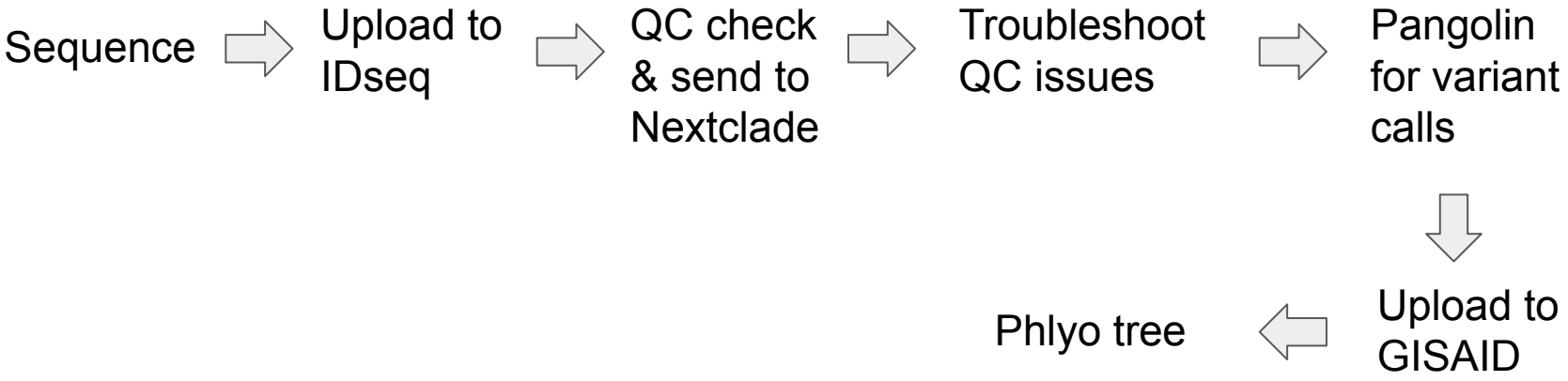
In order to make the trees to interpret transmission, you need to build consensus genomes



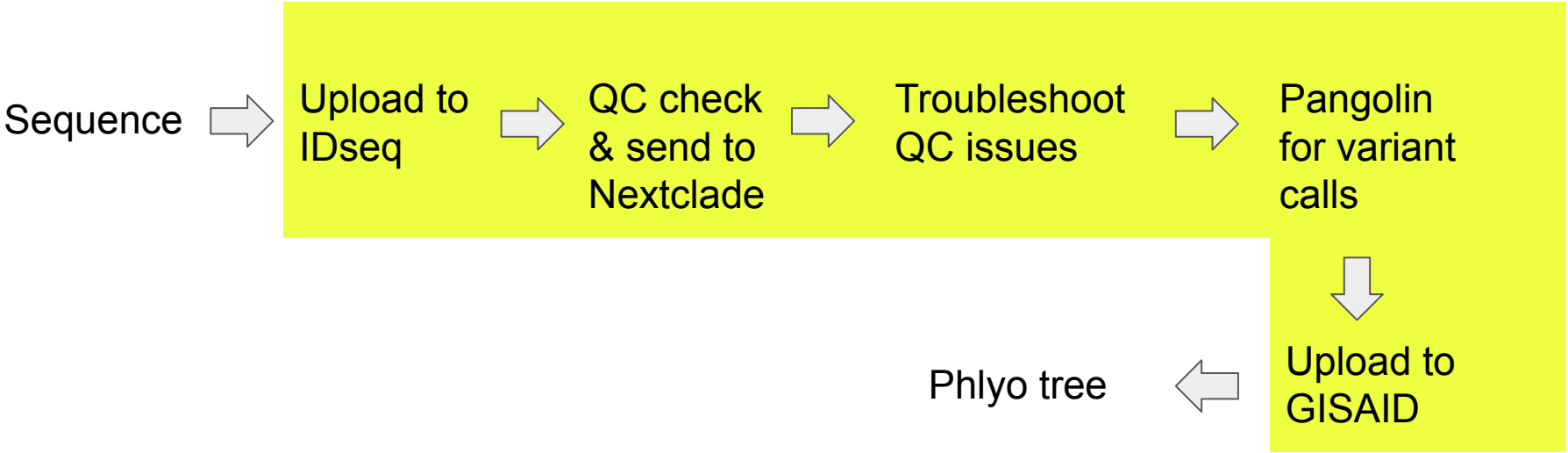
Consensus genome- represents multiple aligned reads

reads	ATTGGAGATAG
	CTTGCATATAG
	AGTGCAGATAGC
	ATTGCAGATAGG
✨ Reference	ATTGCTGATAGG
Consensus	ATTGCAGATAGN

Workflow overview

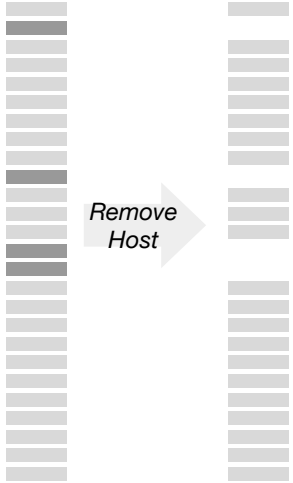


Workflow overview

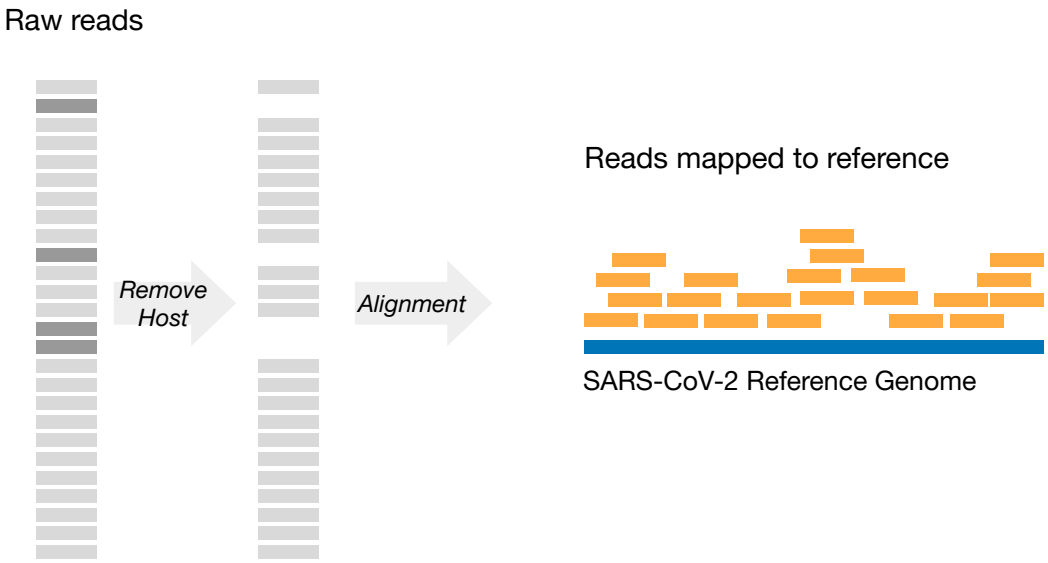


Generating Consensus Genomes

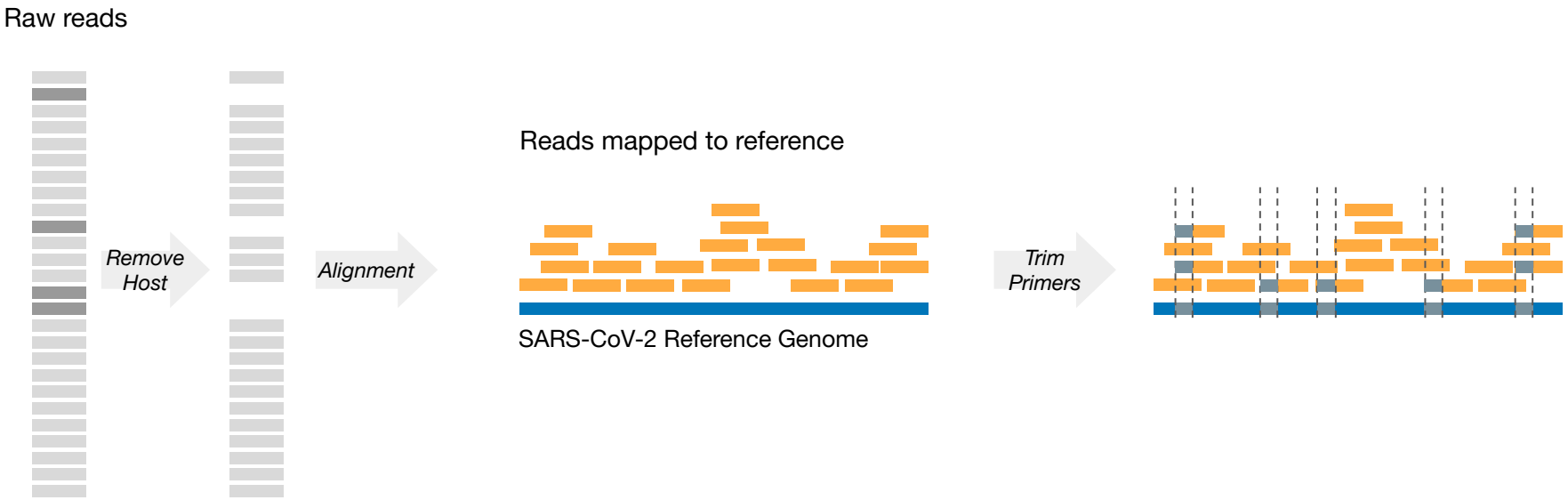
Raw reads



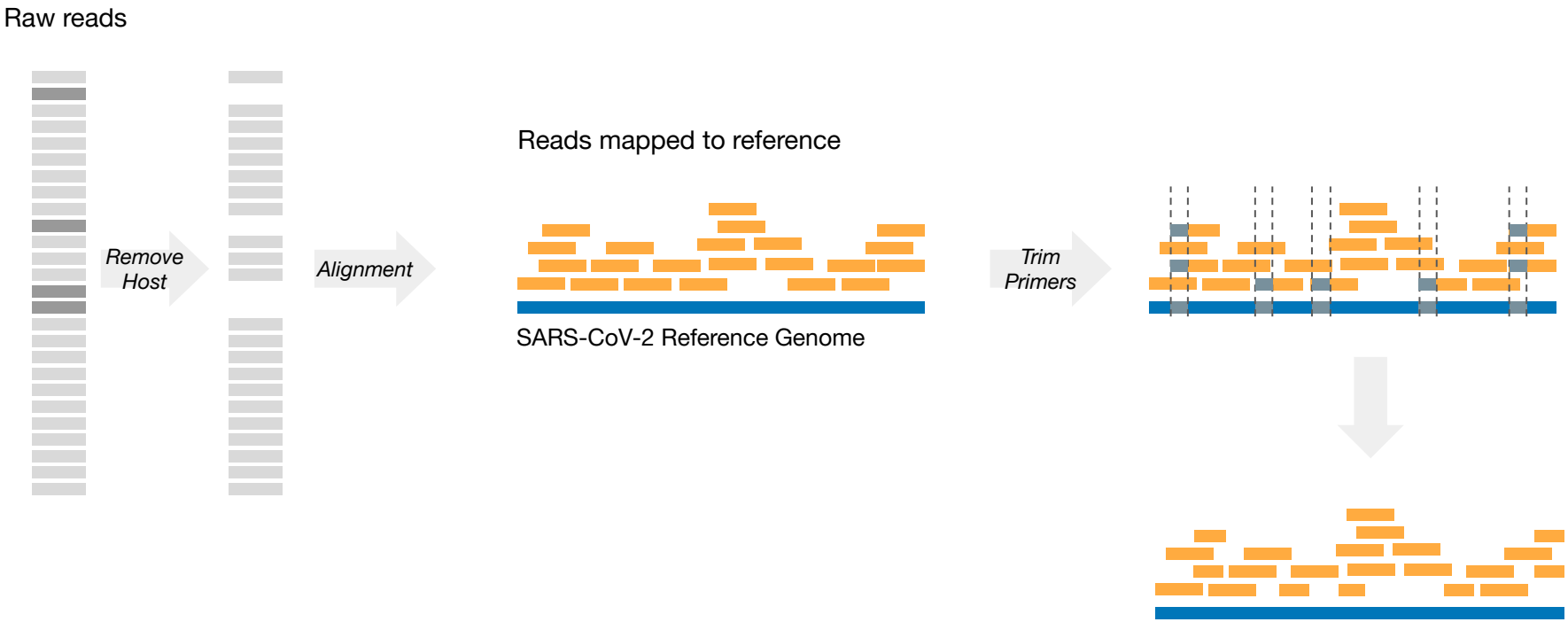
Generating Consensus Genomes



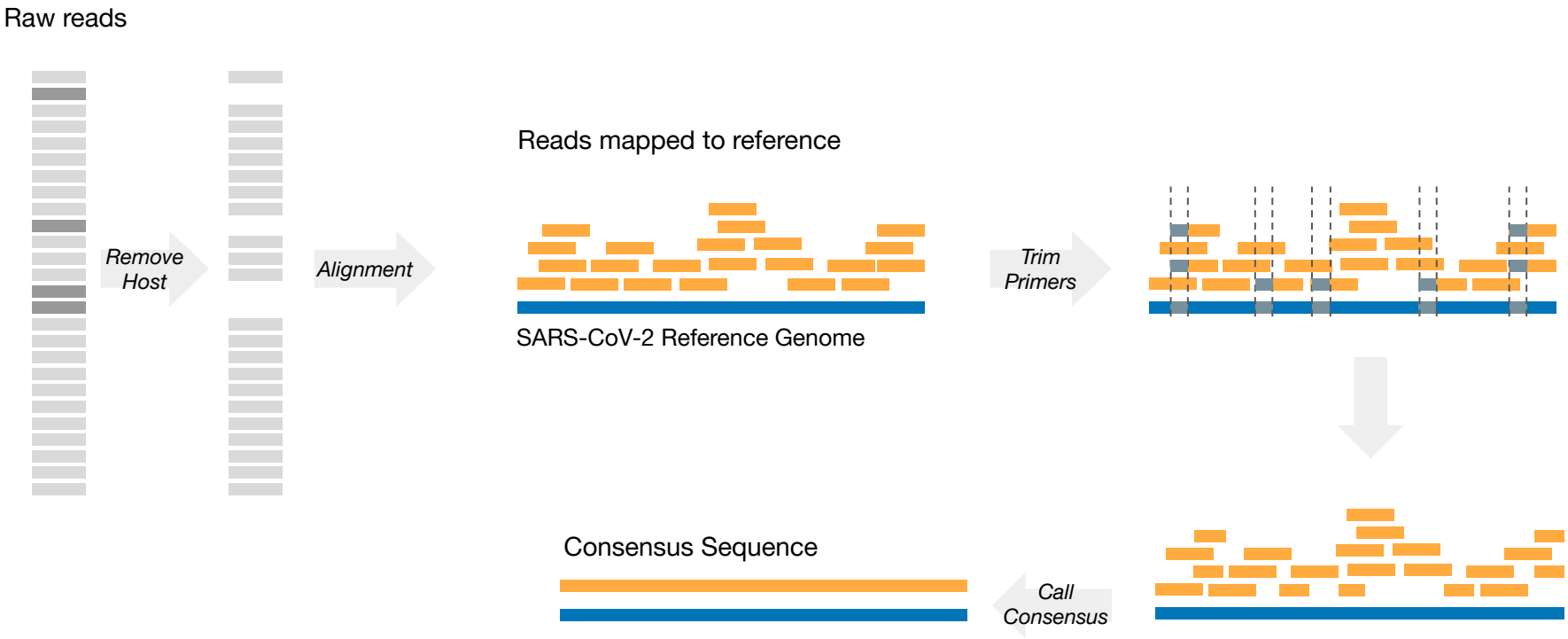
Generating Consensus Genomes



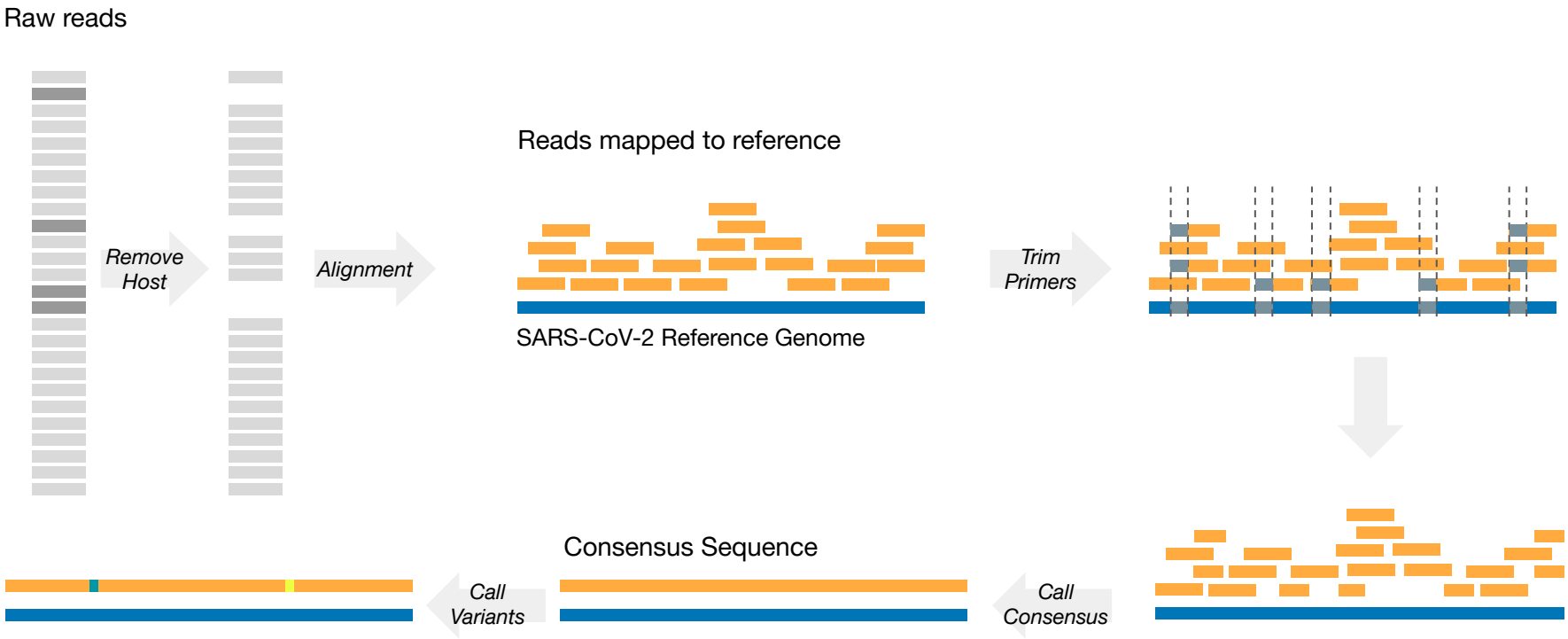
Generating Consensus Genomes



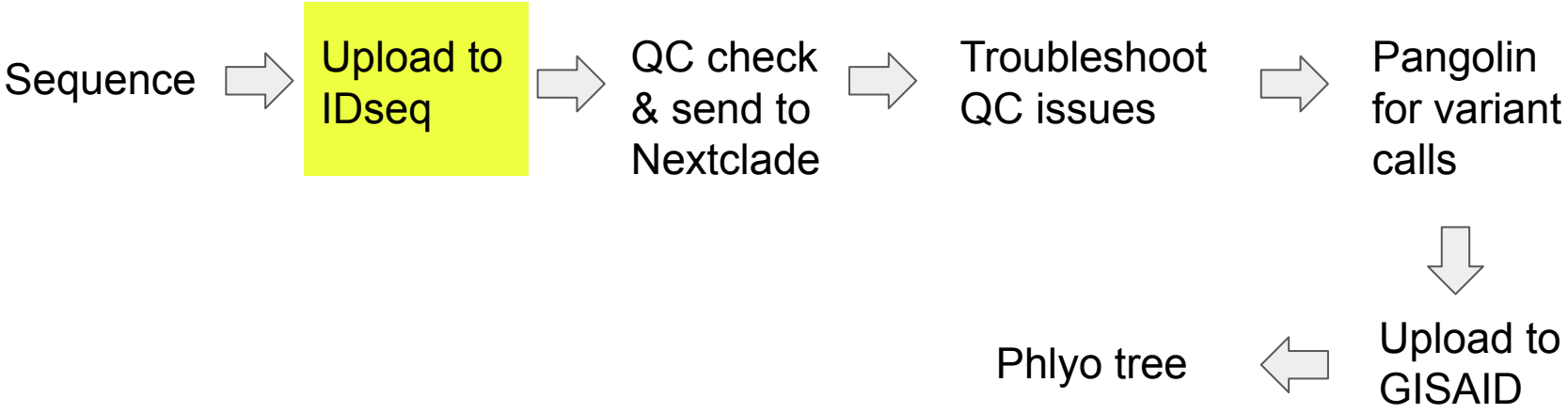
Generating Consensus Genomes



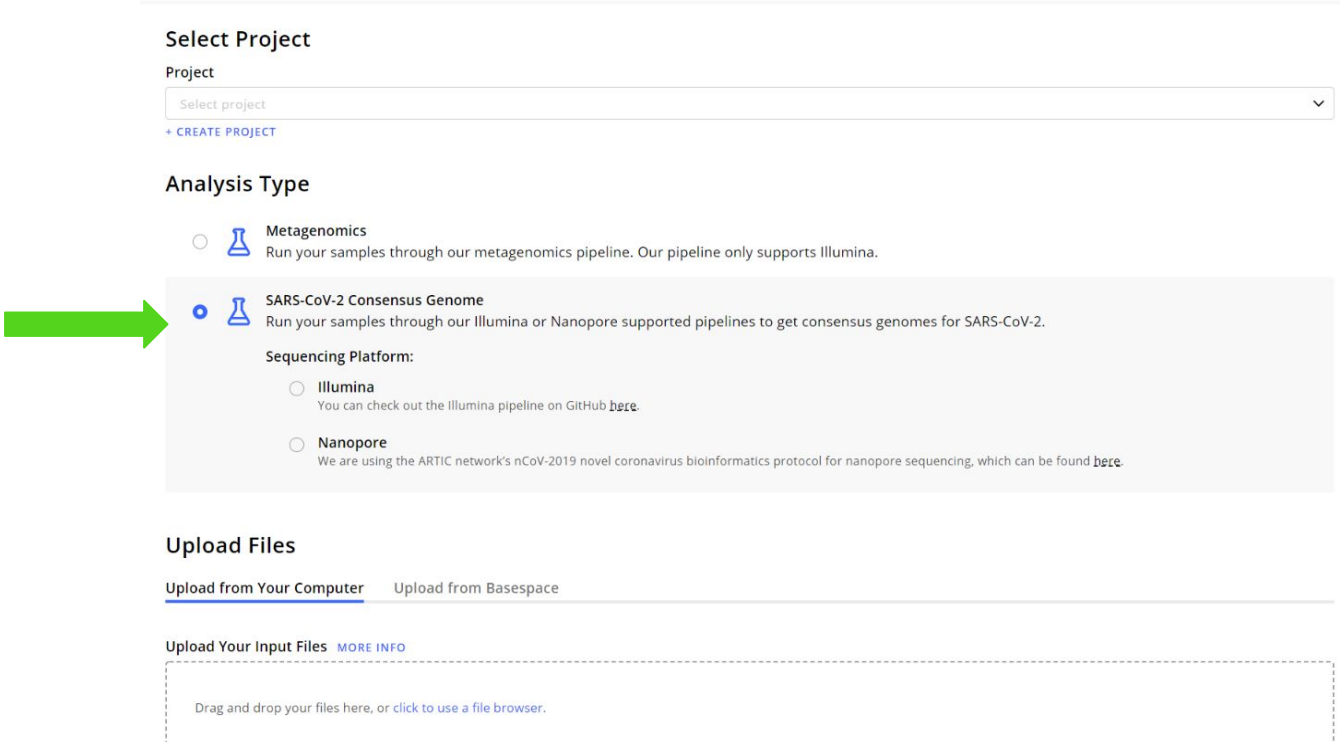
Generating Consensus Genomes



Workflow overview



Supports Illumina and Nanopore platforms



Add metadata

Upload Metadata

This metadata will provide context around your samples and results in IDseq.

1 Samples | 2 Metadata | 3 Review

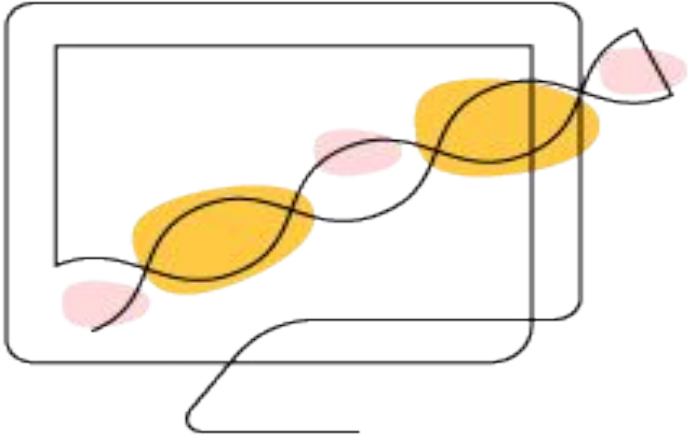
Required fields: We require the following metadata to determine how to process your data and display the results: Host Organism, Sample Type, Water Control, Nucleotide Type, Collection Date, Collection Location. Please be as accurate as possible! View Full Metadata Dictionary.

Available organisms for host subtraction: Human, Mosquito, Tick, Mouse, Cat, Pig, C.elegans, Carp, Chicken, Bee, Salpingoeca rosetta, Bat, Rat, Field Vole, Bank Vole, Rabbit, Water Buffalo, Horse, Taurine Cattle, Turkey, Barred Hamlet, Orange Clownfish, Tiger Tail Seahorse, Torafugu, Avian, White Shrimp.

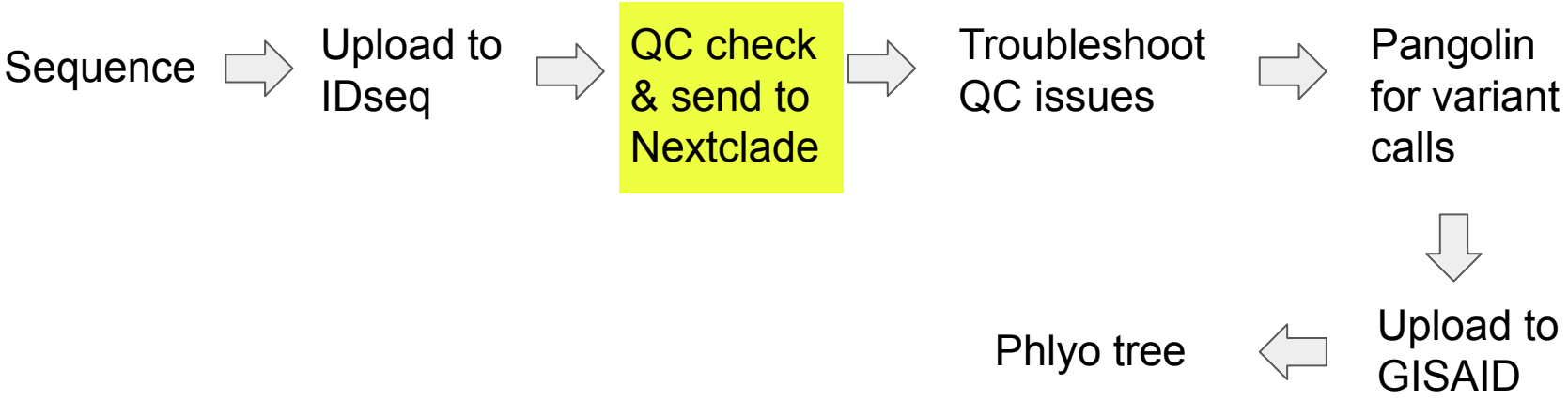
Manual Input | CSV Upload

Sample Name	Host Organism	Sample Type	Water Control	Nucleotide Type	Collection Date	Collection Location
upload_file			No		YYYY-MM-DD	Enter a city, region or country

Pipeline runs automatically in the cloud



Workflow overview



Quality control check

CI ›

sample1_2

Sample Details

Download All

Consensus Genome BETA

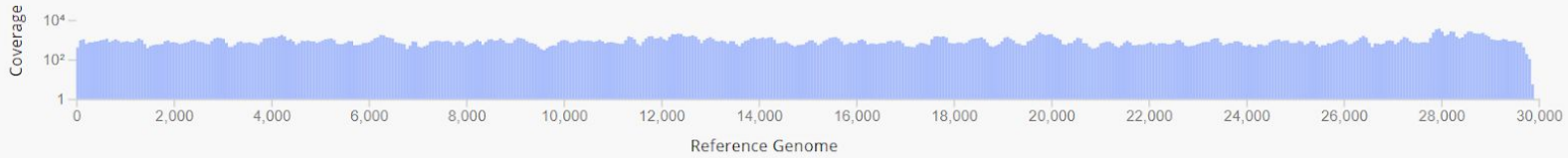
Learn more about consensus genomes ›

Is my consensus genome complete?

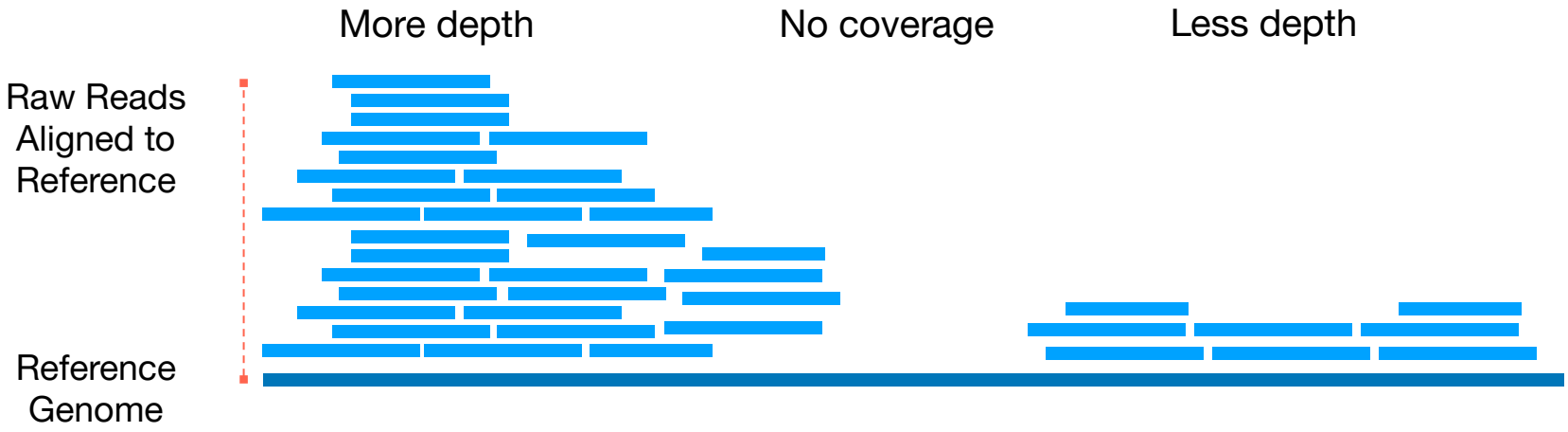
Taxon	Reads	GC Content	SNPs	%id	Informative Nucleotides	Missing Bases	Ambiguous Bases
Severe acute respiratory syndrome coronavirus 2	187444	38.01%	7	100%	29850	12	0

How good is the coverage?

Reference NCBI Entry	MN985325.1	Reference Length	29903	Coverage Depth	889.3x	Coverage Breadth	99.9%



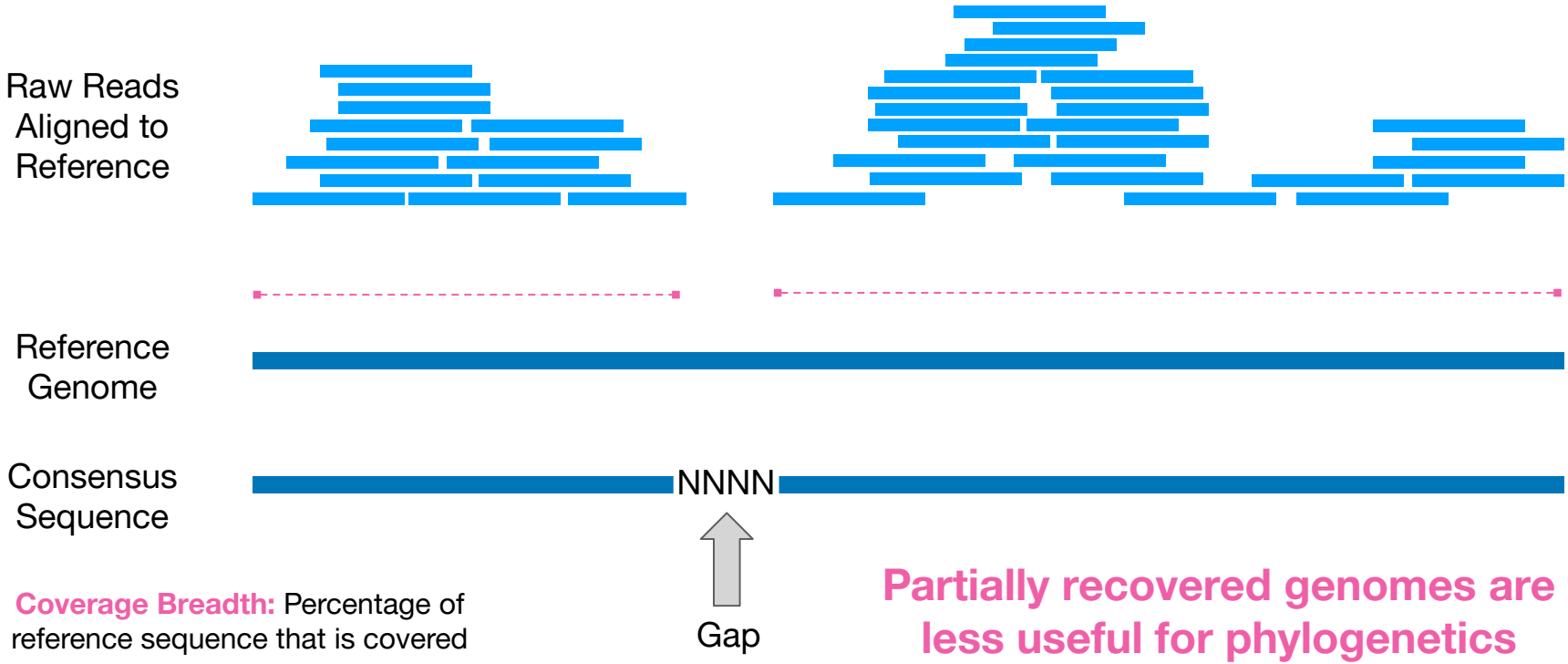
Is there enough depth?



Coverage Depth: # of times a nucleotide is read during sequencing

Must have >10 reads to call a base

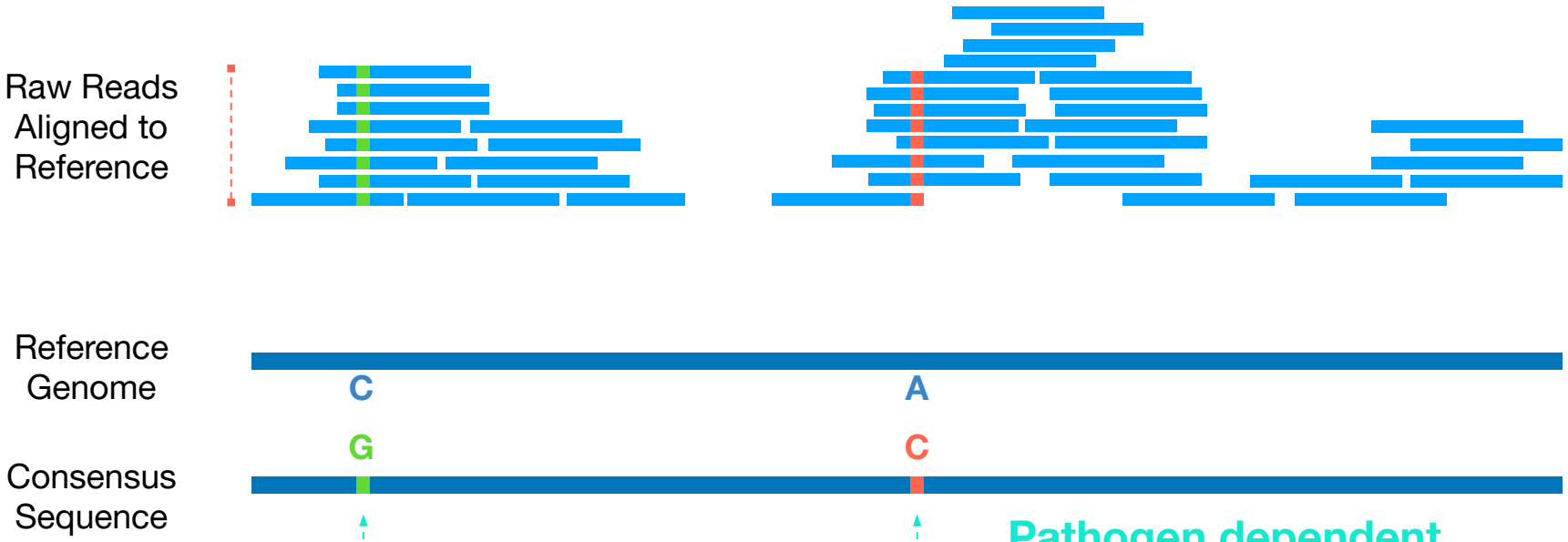
How much of the genome was recovered?



Coverage Breadth: Percentage of reference sequence that is covered

Partially recovered genomes are less useful for phylogenetics

How many SNPs are too many?

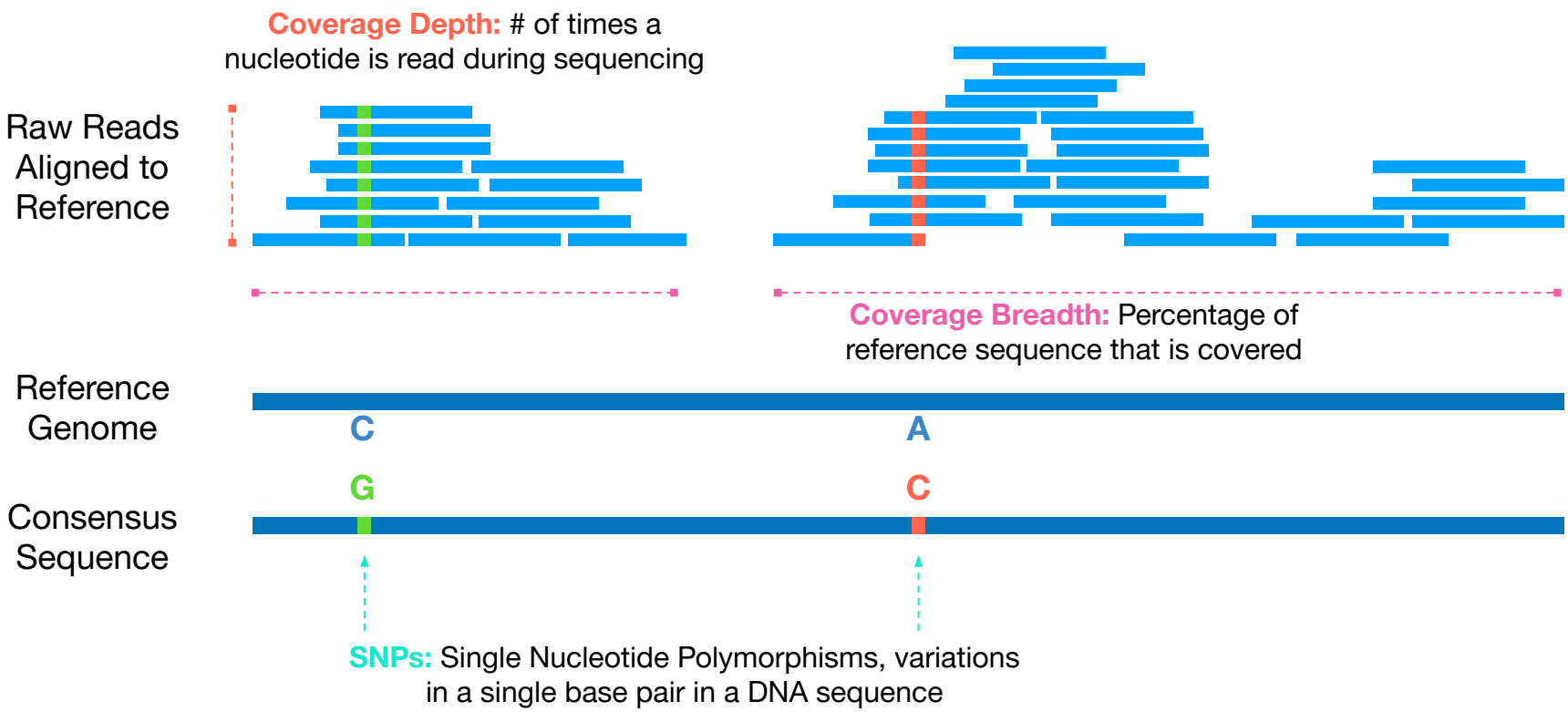


SNPs: Single Nucleotide Polymorphisms, variations in a single base pair in a DNA sequence

Pathogen dependent

- SARS-CoV-2 mutates slowly and shouldn't have too many SNP differences
  - Red flag: 30 differences. Orange flag: 25. Yellow flag: 20.

Evaluating Consensus Genomes

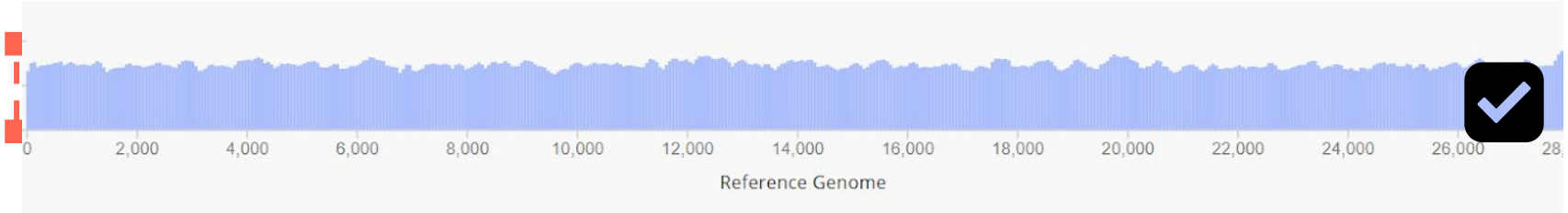


The coverage plot is a great first QC check

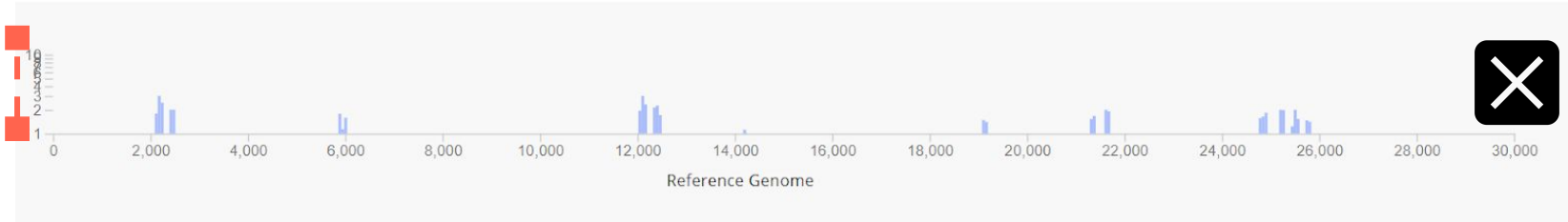
Coverage Depth: # of times a nucleotide is read during sequencing

Must have >10 reads in a location for a base to be called

Raw Reads Aligned to Reference



Raw Reads Aligned to Reference



Important metrics associated with the CG

CI ›

sample1_2

Sample Details

Download All

Consensus Genome BETA

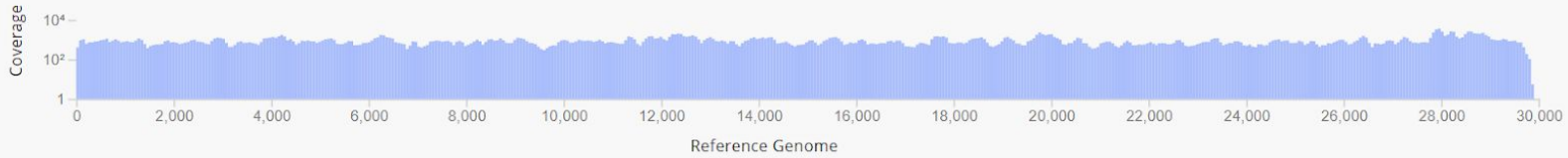
Learn more about consensus genomes ›

Is my consensus genome complete?

Taxon	Reads	GC Content	SNPs	%id	Informative Nucleotides	Missing Bases	Ambiguous Bases
Severe acute respiratory syndrome coronavirus 2	187444	38.01%	7	100%	29850	12	0

How good is the coverage?

Reference NCBI Entry	MN985325.1	Reference Length	29903	Coverage Depth	889.3x	Coverage Breadth	99.9%



Important metrics associated with the CG

CI ›

sample1_2

Sample Details

Download All

Consensus Genome BETA

Learn more about consensus genomes ›

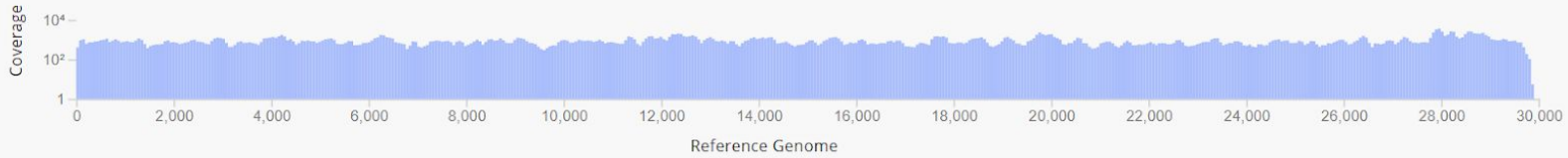
>30

Is my consensus genome complete?

Taxon	Reads	GC Content	SNPs	%id	Informative Nucleotides	Missing Bases	Ambiguous Bases
Severe acute respiratory syndrome coronavirus 2	187444	38.01%	7	100%	29850	12	0

How good is the coverage?

Reference NCBI Entry	MN985325.1	Reference Length	29903	Coverage Depth	889.3x	Coverage Breadth	99.9%



Important metrics associated with the CG

CI ›

sample1_2

Sample Details

Download All

Consensus Genome BETA

Learn more about consensus genomes ›

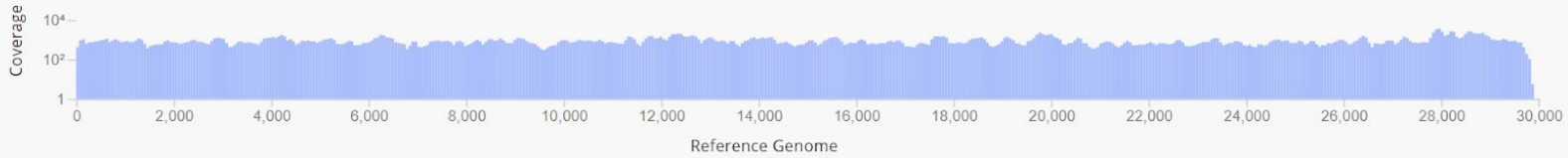
Nextstrain requires 92% of ref genome coverage (>27,510)

Is my consensus genome complete?

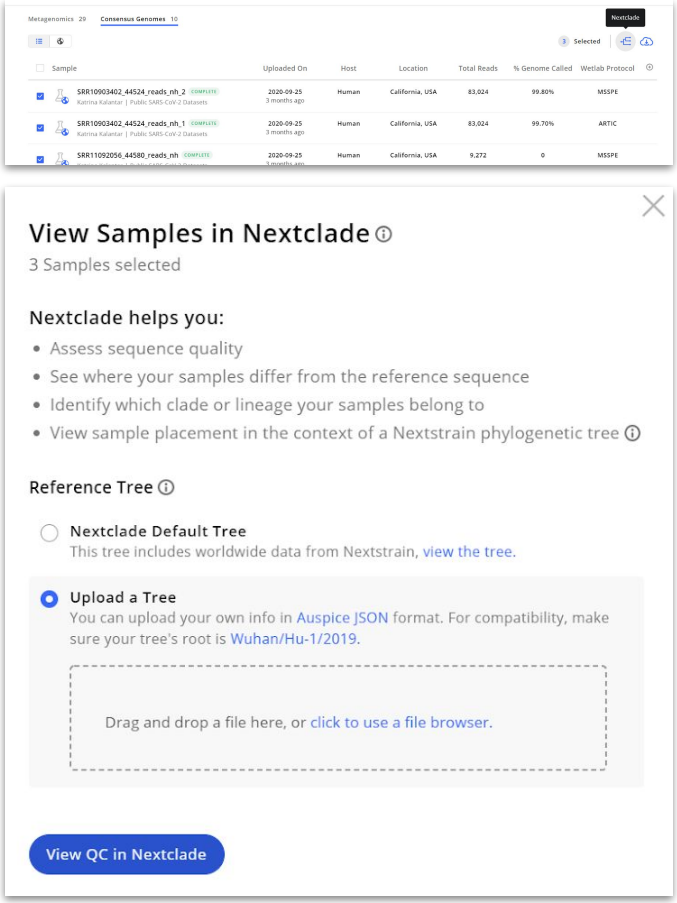
Taxon	Reads	GC Content	SNPs	%id	Informative Nucleotides	Missing Bases	Ambiguous Bases
Severe acute respiratory syndrome coronavirus 2	187444	38.01%	7	100%	29850	12	0

How good is the coverage?

Reference NCBI Entry	MN985325.1	Reference Length	29903	Coverage Depth	889.3x	Coverage Breadth	99.9%



Send samples directly to Nextclade



- Further investigate the quality of your consensus genomes in Nextclade.
- Identify which clade or lineage your sample belongs to.
- Upload an existing tree or use the Nextclade default tree.
- Export auspice.json file from Nextclade.
- View phylogenetic tree with sensitive in a safe and secure environment (Auspice).

Nextclade results

- Sent samples >92% genome coverage (Nextstrain requires this to be added to their builds)

Nextclade

Settings | What's new | English

Back

Done. Total sequences: 8. Succeeded: 8

ID	Sequence name	QC	Clade	Mut.	non-ACGTN	Ns	Gaps	Ins.	Nucleotide sequence
0	CASC20008_L001	N M P C F S	20B	13	1	274	0	0	
1	CASC20011_L001	N M P C F S	20A	14	1	280	0	0	
2	SRR10903402_44524_reads_nh_1	N M P C F S	19A	0	4	60	0	0	
3	SRR10971381_44496_reads_nh	N M P C F S	19A	0	115	3757	0	0	
4	SRR10903402_44524_reads_nh	N M P C F S	19A	0	2	45	0	0	
5	SRR10903401_44525_reads_nh	N M P C F S	19A	0	28	1022	0	0	
6	unknown_S1_L001	N M P C F S	19B	4	6	400	4	0	
7	sample_sars-cov-2_paired	N M P C F S	20C	7	0	12	0	0	

Nextclade: Phylogenetic-based sequence QC

N

Number of sites where a base could not be called: Areas with low or no sequencing coverage have no information to tell you which base should be at that site. These sites are labelled with N's. When a sequence has two many N's it is both hard to align and place on the tree, and thus they are removed from analyses. By default Nextstrain will drop sequencuences with less than 27,000 non ambiguous bases.

M

Mixed sites: If many sequencing reads support *more* than one base at a site, those sites will be designated with an IUPAC ambiguity code, that tells you which *set* of mutations were found at the site. While this can happen given a co-infection event, it more commonly occurs due to sample cross-contamination.

P

Private mutations: If a sequence differs from the Wuhan reference genome by (currently) more than 20 mutations, it will be flagged as having a high number of "private" mutations. The threshold for flagging a sequence as problematic *will be changed* as the diversity of SARS-CoV-2 increases over the pandemic.

Nextclade: Phylogenetic-based sequence QC

C

Clusters of mutations: If your sequence has one or more areas with 6 mutations within a 100nt wide window, then that will be considered a “cluster of mutations” and it will be flagged unless it occurs at a recognized area of the genome. Such clusters of mutations are often artefactual, resulting from challenges aligning the sequence.

S

The presence of premature stop codons: a stop codon within a gene will now result in a QC warning, unless it is one of the very common stop codons in ORF8 at positions 27 or 68. Depending on where it is, it can be the result of an erroneous mutation.

F

The presence of frameshift mutations: This happens when there is an insertion or deletion that causes a gene to have a length that is not divisible by 3. If at least one such gene length is detected, the check is considered "bad". Failure of this check means that the gene likely fails to translate.

Nextclade: Phylogenetic-based sequence QC, in pictures

N

```
CZI    - AGTTCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGTAAC
```

M

```
CZI     - ATGRGAGTAACMGGTARWTTTGACCAGACACACAMGATTBDGGGA
```

P

```
wuhan1  - AGTTGGTCCATGATTCGTTTCGTAAATTCGTCTTCGACAGTTGGT
CZI     - AGTTCGTCCTTGATTGGTTACGTTAATTCGTGTTCGTGAGTTCGT
```

Nextclade: Phylogenetic-based sequence QC, in pictures

C

```
wuhan1  - AGTTGGTCCATGATTCGTTTCGTTAATTCGTCTTCGACAGTTGGT
CZI     - AGTTGGTCCTACGGTGGTTAGAAATTTTCGTGTACCAGAGTTCGT
```

S

```
wuhan1  - AGTTGGTCCATGATTCGTTTCGTCTATTCGTCTTCGACAGTTAA
CZI     - AGTTGGTCCATGATTCGTTTCGTTAATTCGTCTTCGACAGTTAA
```

F

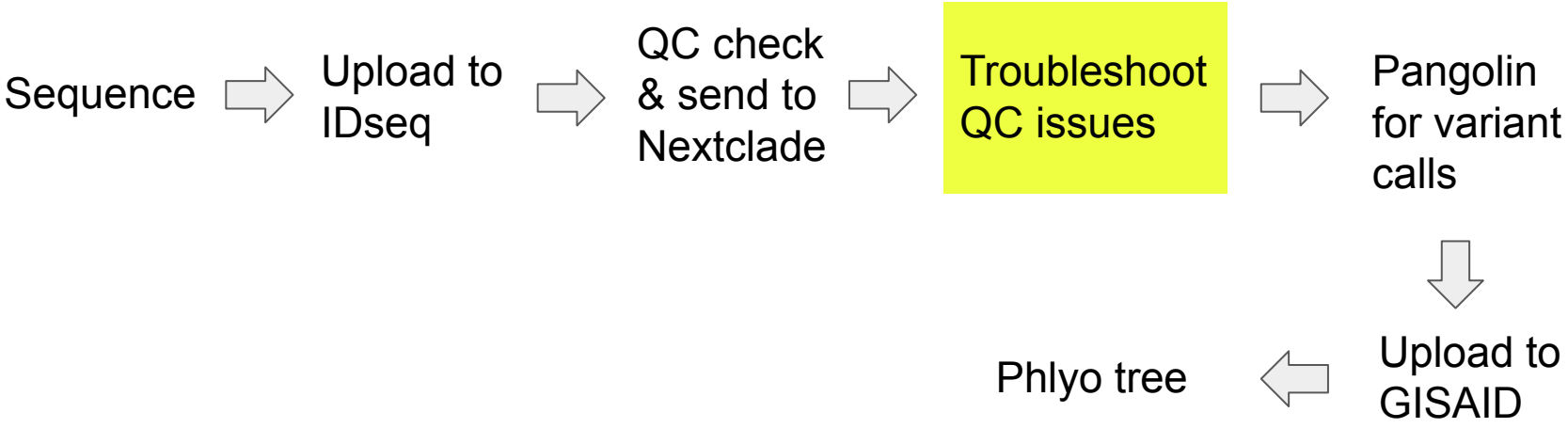
```
wuhan1  - AGT TGG TCC ATG ATT CGT TTC GTC TAT TCG TCT
CZI     - AGT GGT CCA TGA TTC GTT TCG TTA ATT CGT CTTC
```

Other QC checks

Cross contamination

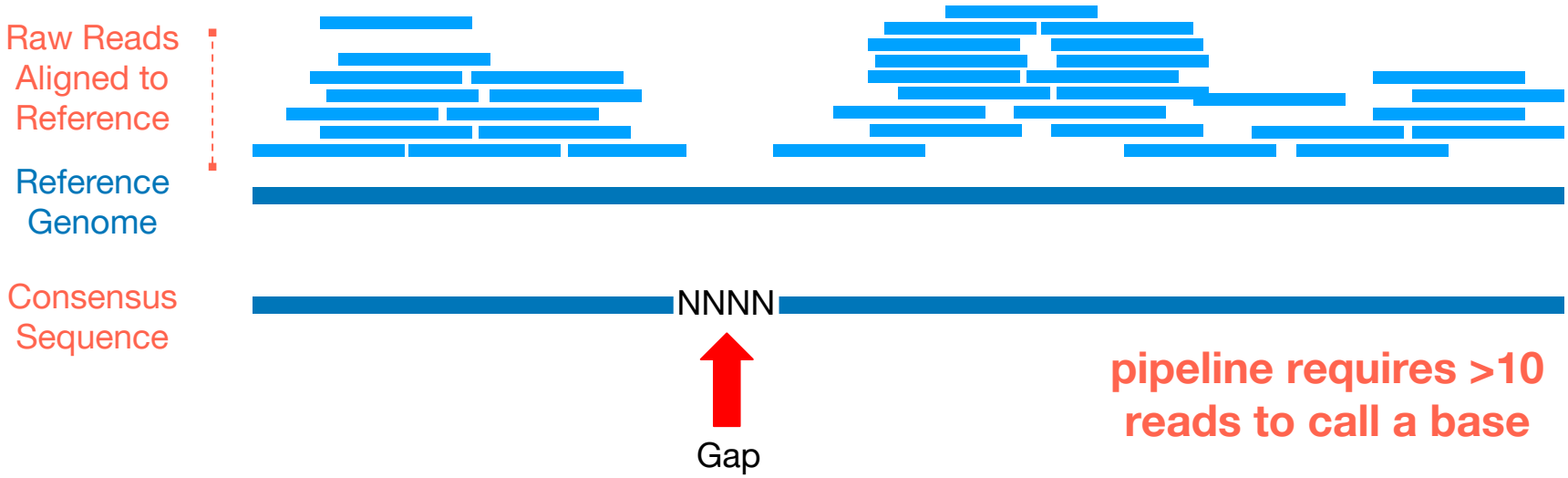
- Always have water controls! Negative controls also good to have
- Normal to see a handful of SARS-CoV-2 reads in controls -- but be concerned if recovering full amplicons, this is a sign of contamination.
- Plate maps -- where are the low Ct samples?

Workflow overview



Troubleshooting too many N's

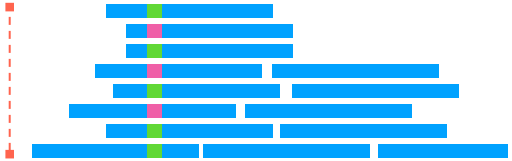
- option to resequence, but should take into account Ct value.
- can concatenate fastq files prior to IDseq upload to double the coverage
- double check sequencing metrics- was this a successful run?



Troubleshooting ‘mixed sites’

- Potential causes: host infected by multiple variants (rare) or contamination
- Contamination check:
  - Check plate map & barcodes used-> shared bardcodes may cause bleedover during sequencing.
- Our pipeline is stringent, can check bam file to see if any bases are confidently called (ie 89% one base and 11% another).
- Make sure to pay attention to where these occur- the ends of reads tend to have lower quality bases

Raw Reads Aligned to Reference



Reference Genome

Consensus Sequence

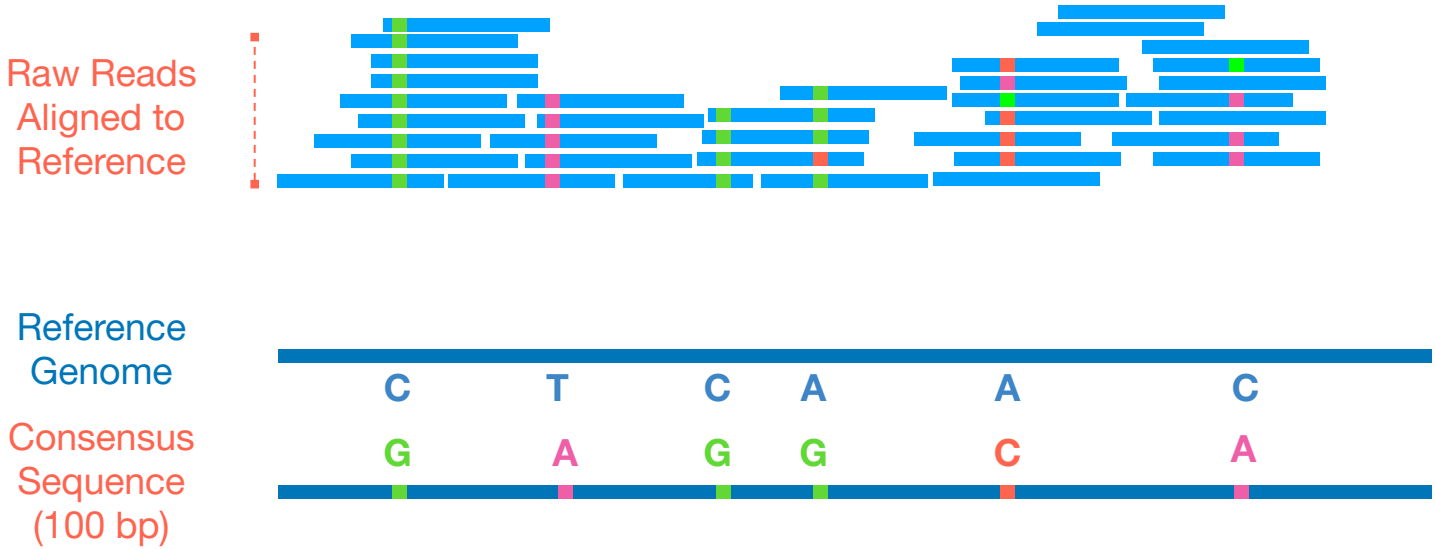
pipeline requires a base to be >90% present to be called

Troubleshooting private mutations

- If there are too many private mutations- viewing the bam file can help.
- What to look for:
  - High coverage in that location all of the reads showing the same base call = good sign it that mutation is real
  - Low coverage and/or reads with different base calls = could be sign of mutations due to contamination

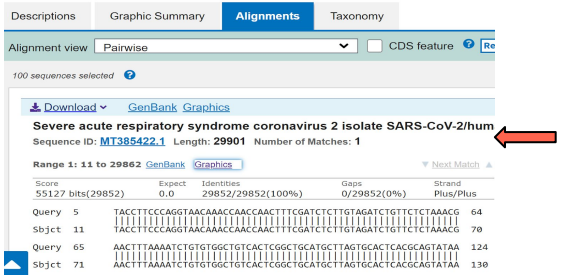
Troubleshooting clusters of mutations

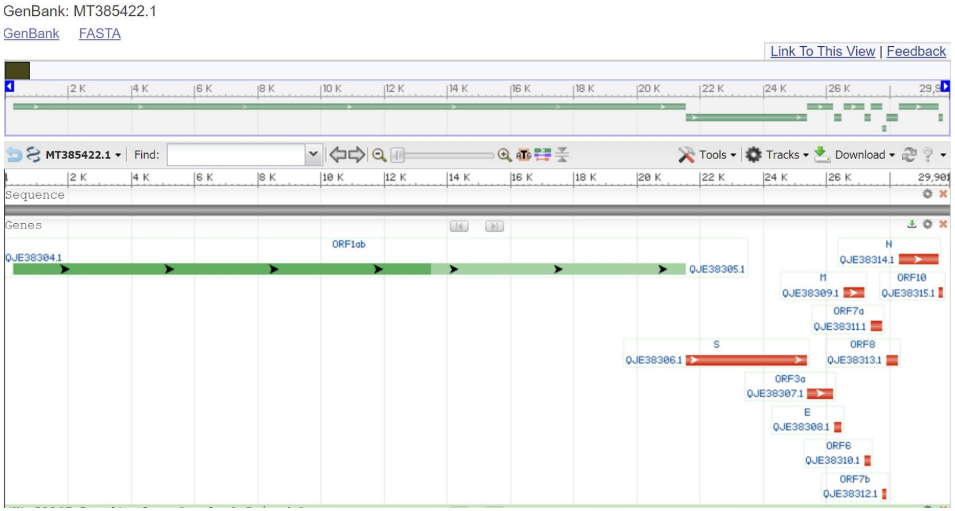
- This usually happens after long stretches of N's



Frameshift mutations

- happen when there are there are deletions or insertions that affect the open reading frames
- Align the consensus genome back to the reference genome
- Check the open reading frames
  - You can do this in BLAST- make sure the ORFs are correct
  - If they are not, have a closer look at the alignment and check out the insertion or deletion.
- Can check bam file
- If there are frameshift mutations the CG won't be accepted to GISAID or Genbank

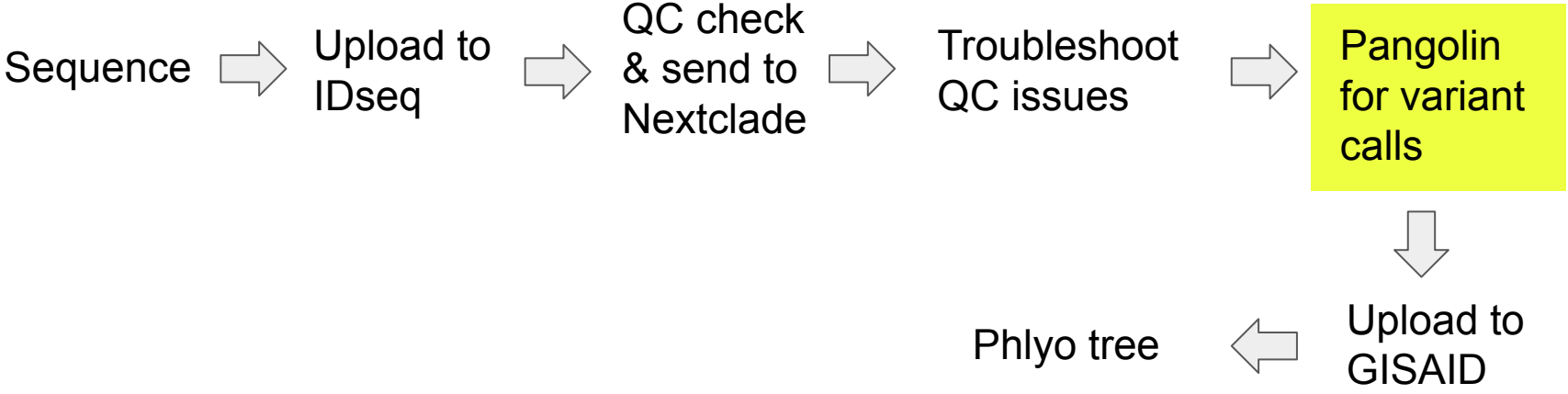




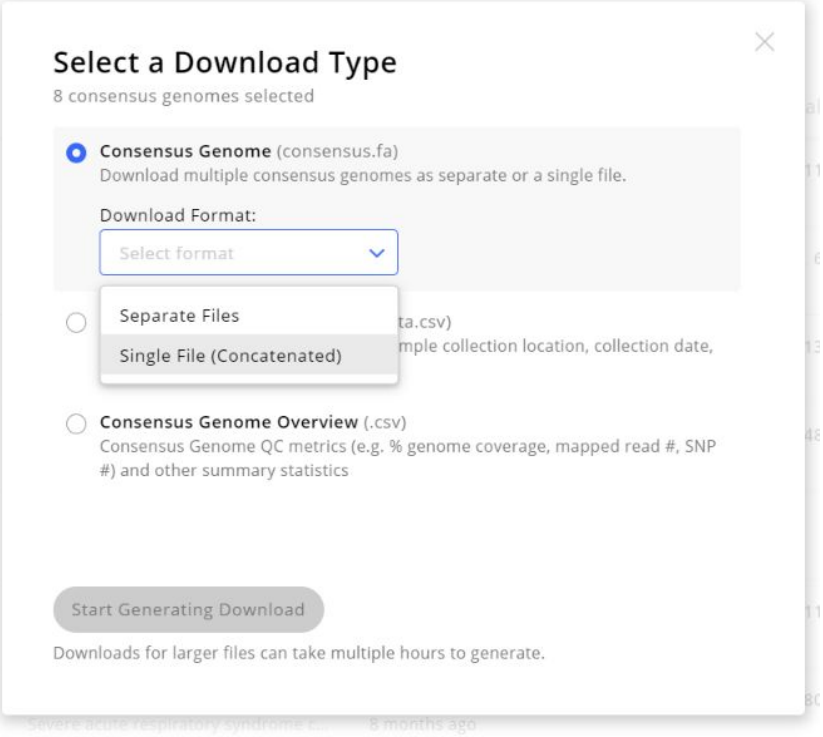
IDseq file outputs and their descriptions

File	Description	Use
consensus.fa	The consensus genome!	The consensus genome
depths.png	Coverage plots	Determine genome coverage
report.tsv	QUAST report	Quality Control
Aligned reads.bam	Initial reads that aligned to the reference genome	Can use in genome browser
ercc_stats.txt	ERCC spike in stats	Used for QC of ERCC control
no_host_1.fq.gz & no_host_2.fq.gz	Non host raw reads	Upload to SRA
Primer trimmed.bam.bai	Aligned reads with trimmed primers (companion to .bam file	used for interrogating coverage results and ensuring quality mappings
Primer trimmed.bam	Aligned reads with trimmed primers	used for interrogating coverage results and ensuring quality mappings
stats.json	QC	Secondary QC if the coverage looks weird

Workflow overview



Download consensus genomes for variant calling and sending to public repositories



Upload consensus genomes to pangolin

https://cov-lineages.org/pangolin.html

Drag and drop fasta file

Select fasta file to upload

Start analysis | Reset entries | Upload another file

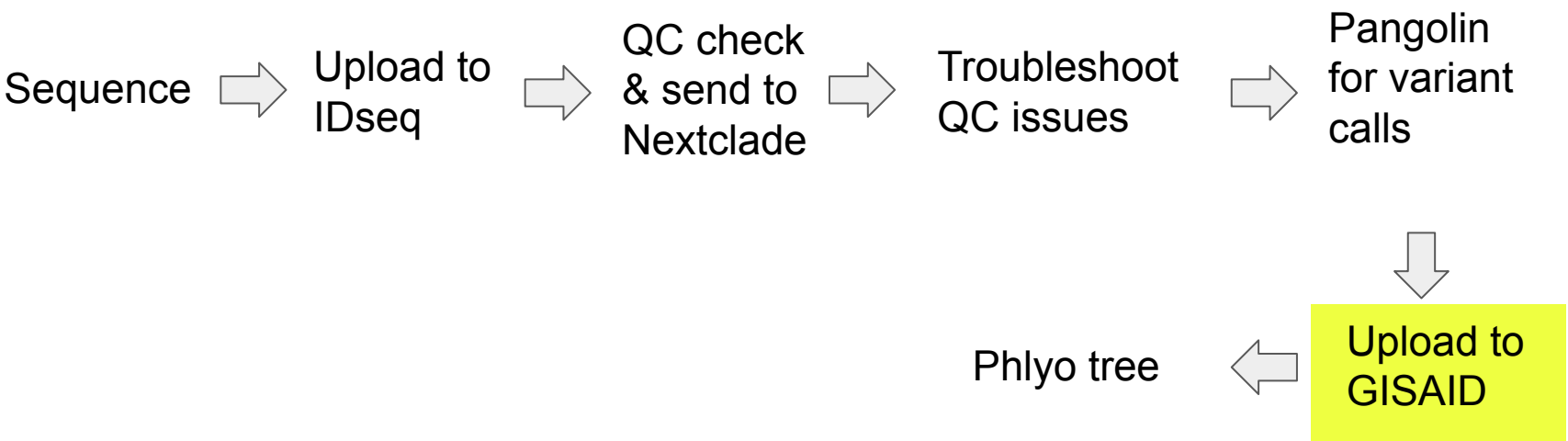
File name	Sequence name
)Y FOR ANALYSIS 8 sequences	
Consensus Genome (3).fa	CASC20008_L(
Consensus Genome (3).fa	CASC20011_L(
Consensus Genome (3).fa	SRR10903402_
Consensus Genome (3).fa	SRR10971381_
Consensus Genome (3).fa	SRR10903402_
Consensus Genome (3).fa	SRR10903401_
Consensus Genome (3).fa	unknown_S1_L
Consensus Genome (3).fa	sample_sars-c(

Pangolin will assign a lineage long with a probability

Reset entries | Upload another file | Help

	File name	Sequence name	Lineage	Assignment Conflict
— ANALYSED (Click tick icon for more info) 8 sequences				
✓	Consensus Genome (3).fa	CASC20008_L001 MN908947.3	B.1.1.205	
✓	Consensus Genome (3).fa	CASC20011_L001 MN908947.3	B.1.403	0.0
✓	Consensus Genome (3).fa	SRR10903402_44524_reads_nh_1 MN908947.3	B	0.0
✓	Consensus Genome (3).fa	SRR10971381_44496_reads_nh MN908947.3	B	0.0
✓	Consensus Genome (3).fa	SRR10903402_44524_reads_nh MN908947.3	B	0.0
✓	Consensus Genome (3).fa	SRR10903401_44525_reads_nh MN908947.3	B	0.0
✓	Consensus Genome (3).fa	unknown_S1_L001 MN908947.3	A	0.0
✓	Consensus Genome (3).fa	sample_sars-cov-2_paired MN908947.3	B.1	

Workflow overview



Submit fasta and metadata to GISAID

[Detailed protocol found here]

Upload options:

1. Single upload
2. Batch upload -> must explicitly request this function

Single Upload

Enter and upload genetic sequence and metadata, available clinical and epidemiological data, geographical as well as species-specific data. Data will be reviewed by a curator prior to release. An email confirmation will be issued upon release.

Virus detail

Virus name*
hCoV-19/Country/Identifier/2020

Accession ID

Type: betacoronavirus

Passage details/history*
Example: Original, Vero

Sample information

Collection date*
Example: 2020-03-27, 2020-03 (collection in March, specific day unknown), 2020 (collection in 2020, month and day unknown)

Location*
Continent / Country / Region

Additional location information
Example: Cave, Live animal market

Host*
Example: Human, Environment, Canine, Manis javanica, Rhinolophus affinis, unknown

Additional host information
Example: Cruise Ship, Convention, Live animal market

Gender*
Example: Male, Female, or unknown

Patient age*
Example: 65, 7 months, or unknown

Patient status*
Example: Hospitalized, Released, Live, Deceased, unknown

Specimen source
Example: Sputum, Alveolar lavage fluid, Oro-pharyngeal swab, Blood, Tracheal swab, Urine, Stool, Cloakal swab, Organ, Feces, Other

Outbreak Detail
Example: Date, Place, Family cluster

Last vaccinated
provide details if applicable

Treatment
Example: Include drug name, dosage

Submit fasta file with high quality consensus genomes

```
>hCoV-19/USA/CA-CZB-32182/2021
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGATCTGTTC
TCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACG
CAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCT
GCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTTGTC
CGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACAC
GTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGAGAC
TCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGGCTTA
GTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGT
TCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAA
GGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAA
ATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGC
CATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGATCCT
TATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGAACTC
ATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGCCCT
```

```
>hCoV-19/USA/CA-CZB-32181/2021
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGATCTGTTCTCTAAACGA
ACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAAT
TAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGC
TTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTTGTCCGGGTGTGA
CCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCAACTC
AGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAG
GAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGGCTTAGTAGAAGTT
GAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGATGCT
CGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAG
TACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTG
GCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTAC
GGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGATCCTTATGAAGAT
TTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGAACTCATGCGTGAG
CTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGCCCTGATGGCTAC
CCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTGTCC
GAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCAT
GAAATTGCTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAA
ATTAAATTGGCAAAGAAATTTGACATCTTCAATGGGGAATGTCCAAATTTTGTATTTCCC
```

Submission files

Metadata (collected during genome upload)

mandatory/optional	
Submitter	GISAID-Username
FASTA filename	the filename that contains the sequence without path
Virus name	hCoV-19/USA/CA-CZB-01/2020 (Must match name in fasta file)
Type	"betacoronavirus" (fixed)
Passage details/history	“Original” (fixed)
Collection date	
Location	North America / USA / California / Contra Costa County
Additional location information	e.g. Cruise Ship, Convention, Live animal market
Host	“Human” (fixed)
Additional host information	e.g. Patient infected while traveling in ….
Sampling Strategy	e.g. Sentinel surveillance (ILI), Sentinel surveillance (ARI), Sentinel surveillance (SARI), Non-sentinel-surveillance (hospital), Non-sentinel-surveillance (GP network), Longitudinal sampling on same patient(s), S gene dropout
Gender	Male, Female, or *unknown*
Patient age	e.g. 65 or 7 months, or *unknown*
Patient status	e.g. Hospitalized, Released, Live, Deceased, or *unknown*
Specimen source	Nasopharyngeal/oropharyngeal swab
Outbreak	Date, Location e.g. type of gathering, Family cluster, etc.
Last vaccinated	provide details if applicable
Treatment	Include drug name, dosage
Sequencing technology	Illumina Miseq
Assembly method	minimap2 / iVar
Coverage	e.g. 70x, 1,000x, 10,000x (average)
Originating lab	Where the clinical specimen or virus isolate was first obtained
Address	
Sample ID given by the originating laboratory	
Submitting lab	Where sequence data have been generated and submitted to GISAID
Address	
Sample ID given by the submitting laboratory	
Authors	a comma separated list of Authors with complete First followed by Last Name

GISAID

You are logged in as **Dan Lu** - logout

Registered Users | EpiFlu™ | **EpiCoV™** | My profile

EpiCoV™ | **Search** | **Downloads** | **Upload** | **My Unreleased**

GISAID hCoV-19 Batch Upload

Upload genetic sequence as single FASTA-File and metadata, available clinical and epidemiological data, geographical as well as species-specific data as XLS or CSV. Data will be reviewed by a curator prior to release. An email confirmation will be issued upon release.

Metadata as Excel or CSV*

max size: 5M Choose File No file chosen

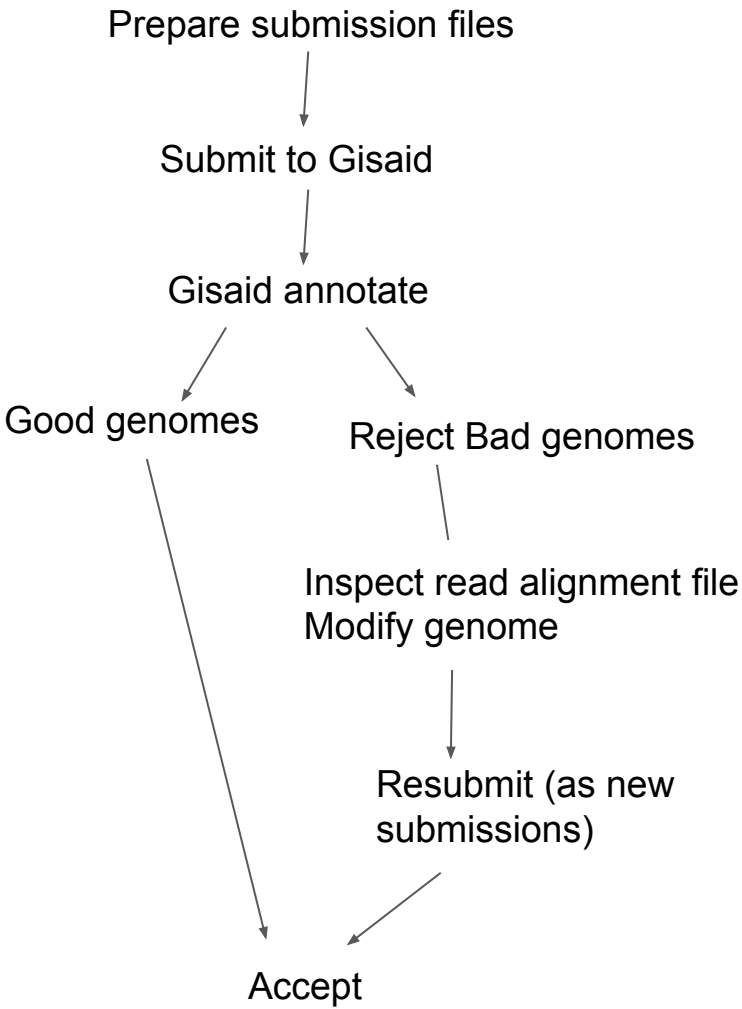
Sequences as FASTA*

max size: 32M Choose File No file chosen

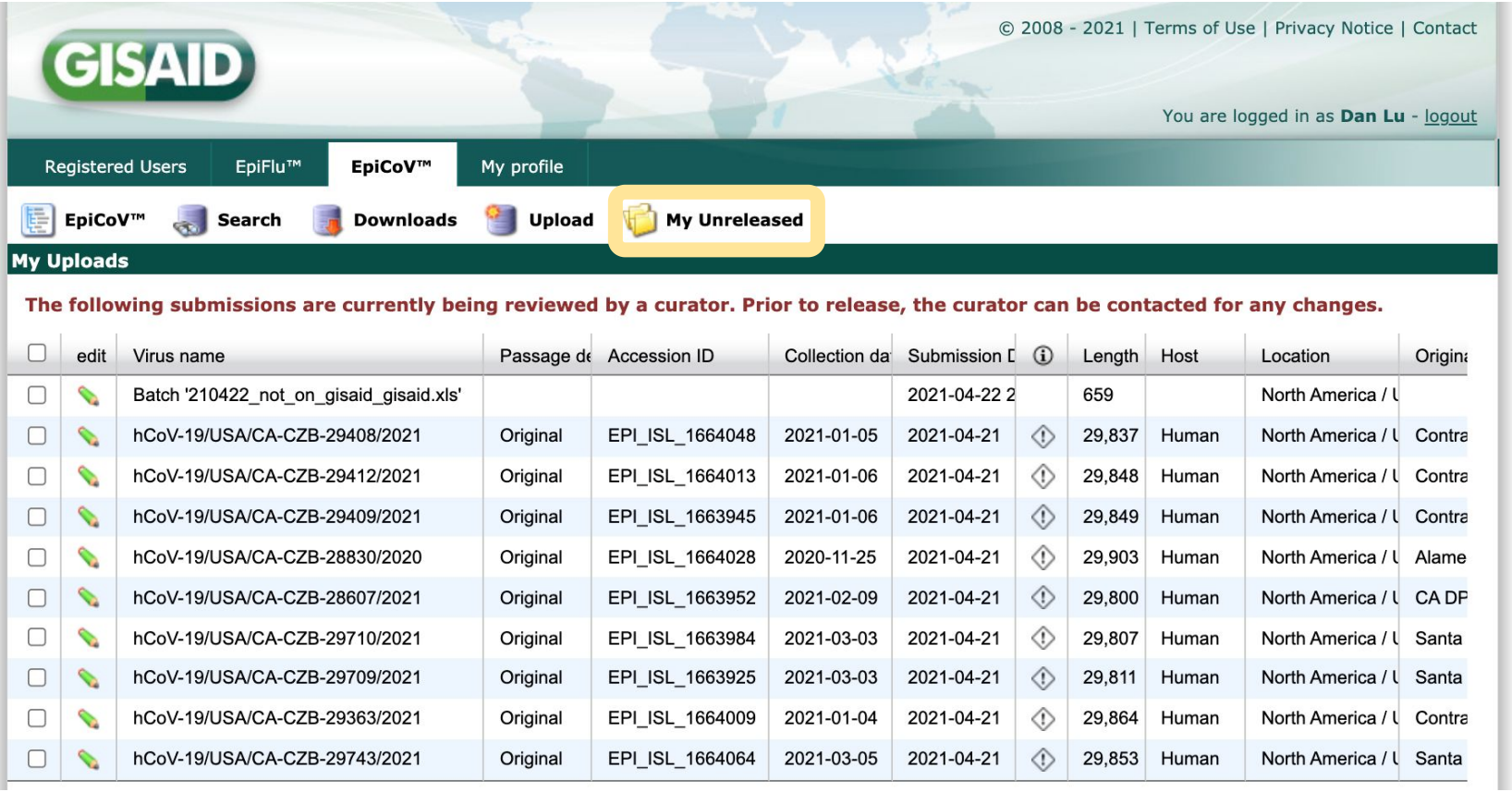
Report

```
Upload XLS/CSV and FASTA.
```

Download Instructions and Template | Contact Curator | Check and Submit


Prepare submission files
Submit to Gisaid
Gisaid annotate
Good genomes
Reject Bad genomes
Inspect read alignment file
Modify genome
Resubmit (as new submissions)
Accept

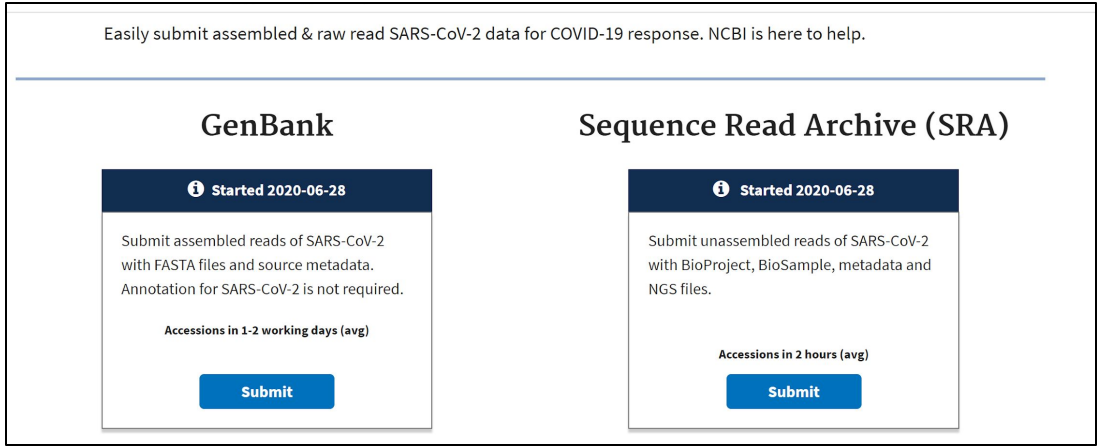
Rejected sequences will have Accession ID assigned, and resubmission of modified genomes needs to go through either a curator or the website



	edit	Virus name	Passage de	Accession ID	Collection da	Submission D		Length	Host	Location	Origina
		Batch '210422_not_on_gisaid_gisaid.xls'				2021-04-22 2		659		North America / U	
		hCoV-19/USA/CA-CZB-29408/2021	Original	EPI_ISL_1664048	2021-01-05	2021-04-21		29,837	Human	North America / U	Contra
		hCoV-19/USA/CA-CZB-29412/2021	Original	EPI_ISL_1664013	2021-01-06	2021-04-21		29,848	Human	North America / U	Contra
		hCoV-19/USA/CA-CZB-29409/2021	Original	EPI_ISL_1663945	2021-01-06	2021-04-21		29,849	Human	North America / U	Contra
		hCoV-19/USA/CA-CZB-28830/2020	Original	EPI_ISL_1664028	2020-11-25	2021-04-21		29,903	Human	North America / U	Alame
		hCoV-19/USA/CA-CZB-28607/2021	Original	EPI_ISL_1663952	2021-02-09	2021-04-21		29,800	Human	North America / U	CA DP
		hCoV-19/USA/CA-CZB-29710/2021	Original	EPI_ISL_1663984	2021-03-03	2021-04-21		29,807	Human	North America / U	Santa
		hCoV-19/USA/CA-CZB-29709/2021	Original	EPI_ISL_1663925	2021-03-03	2021-04-21		29,811	Human	North America / U	Santa
		hCoV-19/USA/CA-CZB-29363/2021	Original	EPI_ISL_1664009	2021-01-04	2021-04-21		29,864	Human	North America / U	Contra
		hCoV-19/USA/CA-CZB-29743/2021	Original	EPI_ISL_1664064	2021-03-05	2021-04-21		29,853	Human	North America / U	Santa

Upload to SRA

- No_host_1.fg.gz & No_host_2.fg.gz
- The raw reads with human reads filtered out
- Upload to SRA



Workflow overview